

Retrieval of Experiments by Efficient Estimation of Marginal Likelihood

Sohan Seth¹, John Shawe-Taylor², Samuel Kaski^{1,3}

¹Helsinki Institute for Information Technology HIIT,
Department of Information of Computer Science, Aalto University, Finland

²Centre for Computational Statistics and Machine Learning,
University College London, UK

³Helsinki Institute for Information Technology HIIT,

Department of Computer Science, University of Helsinki, Finland

¹sohan.seth@hiit.fi, ²j.shawe-taylor@ucl.ac.uk, ³samuel.kaski@hiit.fi

Abstract

We study the task of retrieving relevant experiments given a query experiment. By experiment, we mean a collection of measurements from a set of ‘covariates’ and the associated ‘outcomes’. While similar experiments can be retrieved by comparing available ‘annotations’, this approach ignores the valuable information available in the measurements themselves. To incorporate this information in the retrieval task, we suggest employing a retrieval metric that utilizes probabilistic models learned from the measurements. We argue that such a metric is a sensible measure of similarity between two experiments since it permits inclusion of experiment-specific prior knowledge. However, accurate models are often not analytical, and one must resort to storing posterior samples which demands considerable resources. Therefore, we study strategies to select informative posterior samples to reduce the computational load while maintaining the retrieval performance. We demonstrate the efficacy of our approach on simulated data with simple linear regression as the models, and real world datasets.

1 Introduction

An experiment is an organized procedure for validating a hypothesis, and usually comprises measurements over a set of variables that are either varied (covariates or independent variables) or studied (outcomes or dependent variables). For example, in the study of genome-wide association, one explores the association between ‘traits’ (controlled variable) and common genetic variations (response variables) [1], or in the study of functional genomics covariates can be the species, disease state, and cell type, whereas outcome can be microarray measurements [2].

Traditionally, similar experiments have been retrieved from qualitative assessment of related scientific documents without explicitly handling the experimental data. Recent technological advances have allowed researchers to both acquire measurements in an unprecedented scale throughout the globe, and to release these measurements for public use after curation, e.g., [3]. However, exploring similar experiments still relies on comparing the manual annotations which suffer extensively from variations in terminology, and incompleteness in annotations, e.g., [4]. The global effort of availing researchers with wealth of data invites the need for sophisticated retrieval systems that look beyond annotations in comparing related experiments to improve accessibility.

The next step toward this goal is to compare the *knowledge* acquired from experimental measurements rather than just annotations. From a Bayesian perspective, one can quantify knowledge as the posterior distribution which captures both the information content of the measurements, in terms of the likelihood

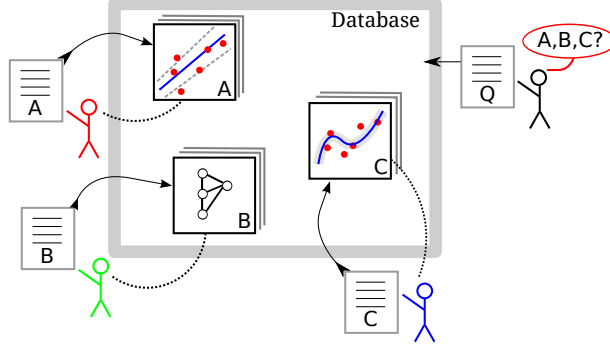


Figure 1: This figure illustrates the basic task we are tackling. Our general objective is to retrieve experiments A, B or C, given query experiment Q. We achieve this by measuring similarity between experiments in terms of the marginal likelihood of the query experiment on the model of the existing experiment. Thus, we assume that the database contains models of experiments learned by the experimenter along with the experimental details. Each model is represented in terms of posterior samples. Our aim is to devise methods to select informative posterior samples to reduce storage and computational requirements while preserving the retrieval accuracy. It is to be noted that we can only compare two experiments if they ‘share’ some common covariates or outcomes.

function, as well as the experience and expertise of the experimenter in terms of the prior distribution. We explicitly assume that we have access to a database where researchers have submitted models learned on the experiment along with measurements and annotations. We study efficient approaches for retrieving relevant experiments utilizing this set-up as a first step toward realizing such an engine.

We suggest the conditional *marginal likelihood* (1) as a similarity metric, where the underlying idea is to evaluate the likelihood of the query experiment on models learned from (individual) existing experiments. Although the suggested metric can be efficiently estimated as the average posterior likelihoods over the posterior samples (2), this approach has two issues: storing the posterior samples requires considerable resource, and evaluating each marginal likelihood can be computationally demanding. This paper deals with selecting *informative* posterior samples to reduce both storage and computational requirements while maintaining the retrieval performance.

We achieve this by approximating the marginal likelihood as a *weighted average* of individual likelihoods over posterior samples (3). The weights are then learned to preserve the relative order of experiments in a training set (section 2.1). This is done while imposing a suitable sparsity constraint which allows us to only consider posterior samples with non-zero weights when computing the likelihood of a query sample, thus reducing the storage and computational burden considerably. Fig. 1 illustrates our general objective.

2 Method

We have a set of experiments $\{\mathcal{E}_d\}_{d=1}^D$. Each experiment is defined as a collection of measurements over covariates and outcomes, i.e., $\mathcal{E}_d = \{(\mathbf{x}_{di}, \mathbf{y}_{di})\}_{i=1}^{n_d}$. We assume that each experiment \mathcal{E}_d has been modeled by a model \mathcal{M}_d , producing a set of posterior MCMC samples $\{\theta_{dk}\}_{k=1}^{m_d}$ from each model. Our general objective is to rank the experiments \mathcal{E}_d —actually the models \mathcal{M}_d in the database—according to their relevance to a new query experiment \mathcal{E}_q which is not in the database.

We suggest retrieving similar experiments in terms of their marginal likelihood,

$$\text{ML}_{q|d} = p(\mathcal{E}_q | \mathcal{E}_d) \quad (1)$$

This metric has been previously discussed in the context of document retrieval where its use is motivated by capturing the user’s intent in terms of the likelihood of a set of keywords \mathcal{E}_q being generated by a document

\mathcal{E}_d [5]. In the context of document retrieval the marginal likelihood is usually computed by jointly modeling multiple documents. However, we cannot evaluate this metric by modeling multiple experiments jointly, since we explicitly allow experimenters to submit their models. Therefore, we utilize individual models $p(\cdot|\mathcal{E}_d) \propto p(\mathcal{E}_d|\cdot)\pi_d(\cdot)$ to evaluate the marginal likelihood as $\text{ML}_{q|d} = \mathbb{E}_{p(\cdot|\mathcal{E}_d)}p(\mathcal{E}_q|\cdot)$, where π_d is the prior information specific to experiment d .

The likelihood can be approximated using posterior samples $\{\theta_{dk}\}_{k=1}^{m_d} \sim p(\cdot|\mathcal{E}_d)$ as

$$\widehat{\text{ML}}_{q|d} \approx \frac{1}{m_d} \sum_{k=1}^{m_d} p(\mathcal{E}_q|\theta_{dk}). \quad (2)$$

This approach is computationally demanding since one needs to store multiple posterior samples $\{\theta_{dk}\}$ and evaluate the corresponding likelihoods $p(\mathcal{E}_q|\theta_{dk})$. The technical contribution of this paper is to address this issue by selecting *fewer* posterior samples that are essential in the retrieval task, i.e., discriminative between experiments. Fig. 2 illustrates our technical objective.

We achieve this by approximating the marginal likelihood as

$$\widetilde{\text{ML}}_{q|d} \approx \frac{1}{m_d} \sum_{k=1}^{m_d} w_{dk} \prod_{i=1}^{n_d} p(\mathbf{x}_{qi}, \mathbf{y}_{qi}|\theta_{dk}) \quad (3)$$

where $\mathbf{w}_d = [w_{d1}, \dots, w_{dm_d}]$ is a vector of *sparse non-negative weights*. In this way, the posterior samples for which the corresponding weights are zero can be safely ignored. Since we are effectively estimating the *weighted mean* of a set of values, ideally speaking, \mathbf{w}_d should be a *stochastic* vector: positive values that sum to one. However, we observe that even without explicitly imposing this constraint we can achieve favorable performance, and this simplifies the optimization problem considerably.

2.1 Preserving ranking of experiments

To learn the weights for each experiment, we adapt the concept of *learning to rank* which is a well explored research problem in information retrieval [6]. However, while this approach is usually applied for learning a function over document-query pairs, we utilize the concept in learning weights over posterior samples for all experiments (“documents”) together.

Assume, without loss of generality, that given a query q and two experiments i_1 and i_2 in the database, i_1 ranks higher than i_2 , i.e., $\text{ML}_{q|i_1} > \text{ML}_{q|i_2}$. Therefore, while learning the weights \mathbf{w}_{i_1} and \mathbf{w}_{i_2} , we need to ensure that

$$\sum_k w_{i_1 k} p(\mathcal{E}_q|\theta_{i_1 k}) > \sum_k w_{i_2 k} p(\mathcal{E}_q|\theta_{i_2 k}).$$

When each experiment in the training set is used as a query q , preserving the relative ranks of each pair $\{i_1, i_2\} \subset \{1, \dots, D\} \setminus \{q\}$ translates to needing to satisfy $D(D-1)(D-2)$ binary constraints for learning the weight vectors $\mathbf{w}_1, \dots, \mathbf{w}_D$. Fortunately not all of the constraints are usually required since a user is often interested in retrieving only the top (say, top K) experiments rather than all experiments.

Therefore, we reformulate our approach and, given a query q , focus on preserving the order of top K experiments. Given any experiment q we select the K closest experiments, $I_q^K = \{i_{j_1}, \dots, i_{j_K}\}$, and compare them pairwise with the rest of the $(D-2)$ experiments in the database. Intuitively, this preserves the relative orders among the top K experiments I_q^K , and also ensures that these experiments are ranked higher compared to the rest of the $\{1, \dots, D\} \setminus \{q \cup I_q^K\}$ experiments. This reduces the set of constraints to $KD(D-2)$ where $K \ll D$. Notice that it is certainly feasible to choose different K for different queries.

2.2 Optimization problem

Satisfying the binary constraints can be formalized as a classification problem $\{(\mathbf{X}_l, y_l)\}_{l=1}^L$ with a highly sparse design matrix \mathbf{X} of dimension $L \times m$ (as depicted in Fig. 3), with $L = KD(D-2)$ realizations and $m =$

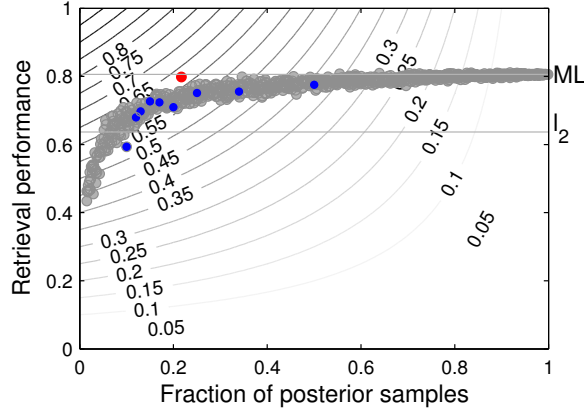


Figure 2: The figure illustrates our objective. To evaluate the marginal likelihood (ML) one can store every k -th posterior sample (blue dots): the choice of k is arbitrary. However, this might not be optimal. For example, selecting posterior samples to be stored randomly (grey dots) might result in better performance. Our goal is to select informative samples from the pool that are discriminative between experiments, to reduce computational requirements without sacrificing retrieval performance (e.g. red dot is achieved by the proposed approach). It is clear that one encounters a trade-off between the sparsity of the posterior samples, and the retrieval performance. Therefore, we utilize $(1\text{-sparsity}) \times \text{retrieval-performance}$ as evaluation metric (contours). In terms of this metric the red dot is close to the best blue and grey dots. ML denotes the performance level when all posterior samples are used. l_2 defines the performance level with l_2 distance based metric.

$\sum_d m_d$ features for learning a combined weight vector $\mathbf{w} = [\mathbf{w}_1, \dots, \mathbf{w}_d]$, i.e., to satisfy $(\mathbf{X}_l \mathbf{w} + b)y_l > 0$ for all l . Each row of \mathbf{X} belongs to a triplet (q, i_1, i_2) , and in that row only the columns associated with posterior samples from i_1 and i_2 are non-zero, and have values $\{p(\mathcal{E}_q|\theta_{i_1k})\}_{k=1}^{m_{i_1}}$ and $\{-p(\mathcal{E}_q|\theta_{i_2k})\}_{k=1}^{m_{i_2}}$ respectively. The label associated with this entry is 1 if $\text{ML}_{q|i_1} > \text{ML}_{q|i_2}$, and zero otherwise. An important aspect of this construction is that the label is not absolute, i.e., we can change the sign of a row in the design matrix, i.e., assign the values $\{-p(\mathcal{E}_q|\theta_{i_1k})\}$ and $\{p(\mathcal{E}_q|\theta_{i_2k})\}$ to the row instead, and switch the label accordingly. Actually, we randomly pick one of these scenarios to maintain class balance, i.e., we have similar numbers of zeros and ones.

Since we are solving a classification problem, each row of the design matrix can be normalized without effecting the class label. This helps solve scaling issues: Instead of likelihoods p_l , we can classify log likelihoods $\ln p_l$, and compute the normalized entries as $\pm \exp(\ln p_l - \max_l \ln p_l)$. These values are in $[-1, 1]$.

We use the library liblinear [7] to solve this optimization problem. We use the logistic cost with l_1 regularization, and set the regularization parameter to 1. An interesting property of this approach is that the number of posterior samples with non-zero weights can be different for different models as needed.

Although we do not restrict the weight vectors to be positive and normalized to one, we observe that the non-negativity becomes satisfied naturally, whereas the sum-up-to-one constraint can be ignored since we are only interested in the ranks. It is to be noted that, since we optimize all the weights together, there is a possibility that all weights from a particular experiment become set to zero to achieve a sparser solution. This can happen in particular when the number of experiments per class is imbalanced, or if an experiment is an outlier in the sense that it is ranked low most of the times. We leave solving this issue, by imposing additional constraints, for future work.

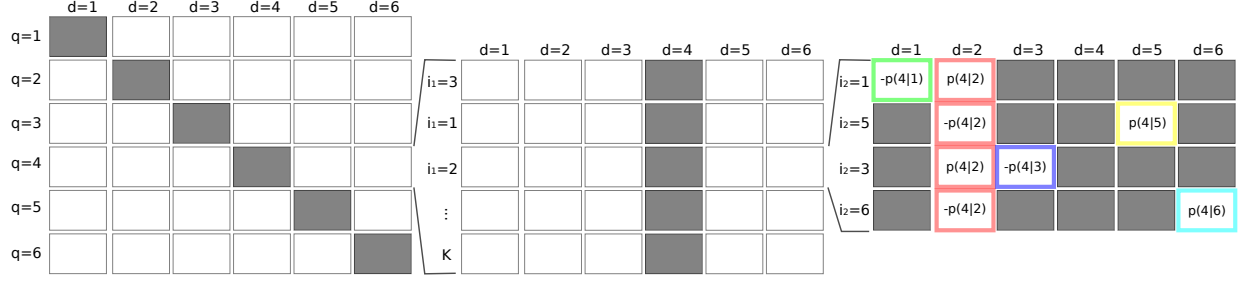


Figure 3: Illustration of design matrix for $D = 6$ with notation $p(a|b) = p(\mathcal{E}_a|\theta_{b,1:m_b})$. The matrix is $D(D - 2)K \times m$ dimensional where $m = \sum_d m_d$ is total number of posterior samples from D experiments. The second and third figure are zoomed versions of a block of rows of the first and second figure respectively. Each row of the design matrix belongs to a triplet (q, i_1, i_2) , a query and two retrieved experiments. The matrix is sparse: each row only has $m_{i_1} + m_{i_2}$ non-zero entries $\{+p(\mathcal{E}_q|\theta_{i_1 1}), \dots, +p(\mathcal{E}_q|\theta_{i_1 m_{i_1}}), -p(\mathcal{E}_q|\theta_{i_2 1}), \dots, -p(\mathcal{E}_q|\theta_{i_2 m_{i_2}})\}$ corresponding to posterior samples of i_1 and i_2 . The signs of the entries and corresponding target can be switched arbitrarily. The matrix contains repeated entries, e.g., the red blocks. We do not consider $ML_{q|q}$, so the diagonal blocks of the first figure are zero (gray).

3 Related works

The state-of-the-art in retrieval of experiments is annotation-driven search, where the user queries with a keyword, and results that match the keyword are returned. For example, the experimental factor ontology [8] provides an excellent platform for retrieving gene expression experiments. However, this approach requires extensive manual curation to fit the different terminologies chosen by different groups and researchers, and is obviously not usable for finding phenomena the experimenter either did not notice or annotate.

One can take a step further, and compare experiments based on the relation between covariates and outcomes, i.e., $f : \mathbf{x} \rightarrow \mathbf{y}$ with some distance metric $d(f_i, f_j)$, where f_i, f_j are *point estimates* of the relations. If f is linear, l_2 can be a suitable distance measure. It is also possible to explore the similarity in either the covariates or the outcomes alone in terms of a suitable representation. This approach has not been taken in the literature yet; it has the obvious limitation of not capturing the uncertainty in f , and possible multimodality of the $p(f)$. We empirically demonstrate that capturing this uncertainty improves the retrieval performance (section 4.1).

We use $ML_{q|d}$ because it implies a very natural definition of relevance: experiment d is relevant to q if a model of q would also be a good model of d . Additionally, the definition brings intrinsic properties not satisfied by most of the other possible approaches. First, one can compare two experiments that do not share the same feature space, for instance, one having missing features; those features can be marginalized out while computing the marginal likelihood. Second, the models for the existing experiments do not need to belong to the same family, and one can choose different models for an experiment as long as the likelihood of the other experiments can be evaluated in terms of that model. Third, since each experiment is modeled separately, the experimenter can include her experiment-specific prior knowledge in the model.

One could also consider the similarity $ML_{d|q} = p(\mathcal{E}_d|\mathcal{E}_q)$ which can be evaluated if one has the model of the query experiment and the measurements from the previous experiments. However, first, this approach would implicitly assume that the querying experimenter is already capable of modeling the experiment properly, which somewhat contradicts the purpose of the retrieval. Second, since experiments \mathcal{E}_d can have different number of observations n_d , this metric is excessively dependent on number of observations: small n_d may result in larger likelihood.

If one models the query experiment as well, then there are other possible approaches of evaluating similarity between two experiments. For example, [9] have recently suggested modeling posterior samples $\{\theta_{dk}\}$ sequentially with Dirichlet process mixtures of normal distributions using particle filtering. Once this

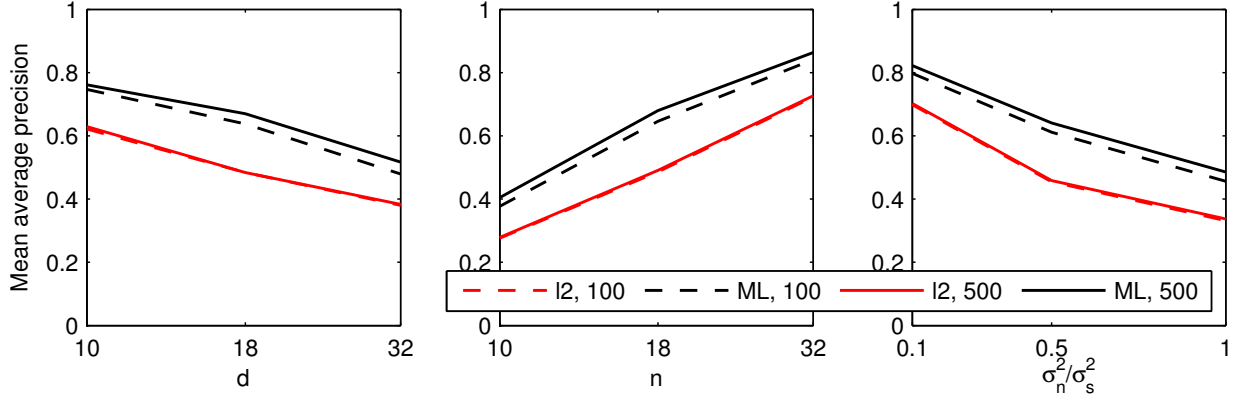


Figure 4: Comparison of retrieval performance of the proposed probabilistic metric $ML_{q|d}$ to an ordinary l_2 metric computed between posterior means of the relation $\hat{f} : \mathbf{x} \rightarrow \mathbf{y}$. Each experiment has been treated as a regression problem with d input features, one output feature, and n measurements, i.e., $f \equiv \mathbf{w}$. The plots show the variation of mean average precision (MAP) as a function of the dimensionality d , number of samples n and signal to noise ratio, for 100 and 500 posterior samples respectively. For each plot the other two parameters have been averaged over. We observe that the probabilistic metric consistently outperforms the l_2 metric. Also the performance of $ML_{q|d}$ improves with the number of posterior samples. See section 4.1 for details.

model (over posterior samples) has been learned, the similarity between two experiments can be evaluated through similarity of the cluster assignments of the respective posterior samples. Given models of the query and the existing experiments, one can also evaluate their similarity in terms of probabilistic distances or kernels [10]. However, both these approaches have the limitation that the models have to belong to the same family for the similarity to be defined. Moreover, the distances or kernels between models are primarily chosen to satisfy only general properties such as the triangular inequality and positive definiteness, rather than assisting in the user’s task, in our case retrieval.

Another possible approach for measuring similarity between experiments is to model the measurements together in a multi-task learning framework [11]. However, off-the-shelf methods for modeling multiple experiments together utilize the same prior and likelihood for all experiments which restricts the generality, and will not exploit the benefit of the knowledge available at the experimenter’s disposal. That said, the true purpose of multi-task learning is to utilize knowledge from similar tasks to improve the learning of a new task, which is fundamentally different than retrieval. Also, treating each experiment or model separately rather than as part of a unified model provides well desired modularity to separate the modeling and retrieval task that can be handled by respective experts.

A similar problem has been explored before by [12] where the authors aimed at retrieving a single sample given a query sample. This was done by modeling multiple samples together using latent Dirichlet allocation. Retrieving an experiment given a query experiment, however, is conceptually very different since a single sample cannot capture the experimental variability that one might be interested in. That said, retrieval of experiments as discussed in this article allows one to also query with a single observation to find the closest experiment which could have generated that particular sample. This approach has an intriguing characteristic that it enables assigning different parts of the query experiment to different models.

4 Simulations

We study the performance and features of the proposed approaches in a simple set-up where the relation between covariates $\mathbf{x} \in \mathbb{R}^d$ and outcome y is assumed to be linear, and corrupted by additive Gaussian noise.

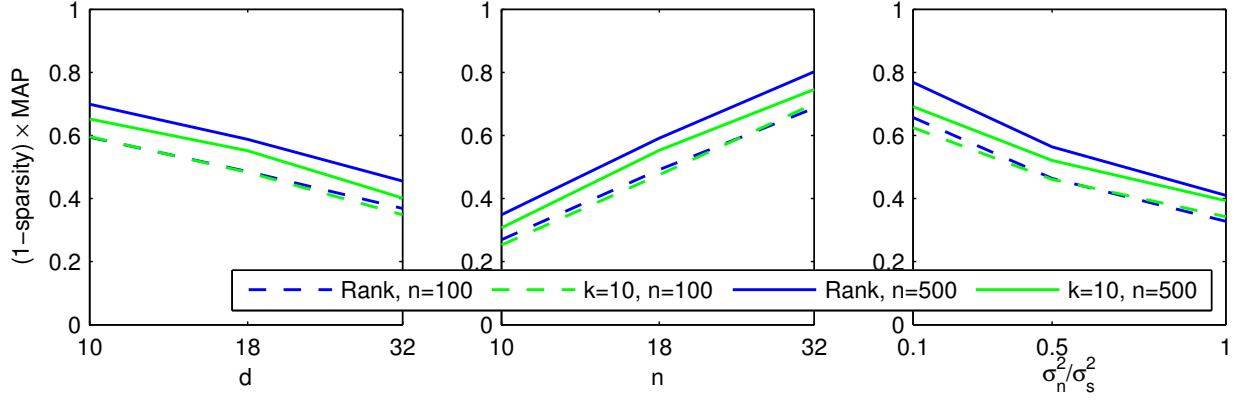


Figure 5: Comparison of retrieval performance between probabilistic metrics estimated from straightforwardly picking every k -th posterior sample, and estimated after learning a weight vector by preserving ranks. The evaluation metric focuses on whether the method has improved retrieval performance while decreasing the number of posterior samples stored. The experiments are simple regression tasks $\mathbf{w} : \mathbb{R}^d \rightarrow \mathbb{R}$ where the ground truth regressors come in clusters. The figures show the performance as a function of the dimensionality (d), number of measurements (n) and signal-to-noise ratio, where the other two features have been averaged over. The total posterior samples are either 100 or 500. We observe that the proposed ranking-based approach outperforms the alternative of storing every k -th posterior sample, in particular for high signal-to-noise-ratios. See section 4.2 for details.

Thus each experiment \mathcal{E}_i can be described by the linear relation $\mathbf{w}_i : y = \mathbf{w}_i^\top \mathbf{x} + \epsilon$. In order to create a *ground truth*, the experiments are assumed to come in clusters, where each cluster is centered at \mathbf{w}_i^* , $i = 1, \dots, C$, where C is the number of clusters. Thus each retrieved experiment can be classified as either relevant or irrelevant depending on whether it shares the same cluster with the query, and the retrieval performance can be evaluated using a standard metric such as *mean average precision* (MAP) [13]. For a fair comparison, we do not use any experiment-specific prior information during modeling since our objective here is to discuss that, first, the proposed retrieval metric performs reasonably well compared to trivial retrieval metrics, and second, rank preservation leads to similar retrieval performance using only a fraction of posterior samples.

We generate experiments $\mathcal{E}_i \equiv \mathbf{w}_i$ from $C = 20$ clusters. The number of experiments within each cluster is generated from a Poisson distribution with rate 10, thus, we have ~ 200 experiments. We randomly split the experiments in two groups: 75% of the experiments are treated as the database and used for training, and the rest are used as queries. The number of measurements in each experiment is chosen from a Poisson distribution with rate n , and we generate m posterior samples from a regression model with Gaussian likelihood and sparse gamma prior over the weight precisions. We use JAGS to generate the posterior samples. To evaluate the performance over different parameter settings we choose $d \in \{10, 18, 32\}$, $n \in \{10, 18, 32\}$, $m \in \{100, 500\}$, and $\sigma_n^2/\sigma_s^2 \in \{0.1, 0.5, 1\}$, thus, 54 set-ups in total.

4.1 Comparison between $\text{ML}_{q|d}$ and $l_2(\hat{\mathbf{w}}_d, \hat{\mathbf{w}}_q)$

We start by comparing the performance of the marginal likelihood metric over the straightforward l_2 metric (Fig. 4). Marginal likelihood consistently outperforms the ordinary distance between posterior means $\hat{\mathbf{w}}_i$. Here we have used all posterior samples for computing $\text{ML}_{q|d}$. Notice that we can easily consider a multi-modal posterior distribution where the posterior mean is not a sufficient descriptor, and this would result in poor retrieval performance for the alternative method. Our goal here was to show that even in simple cases, learning the posterior distribution can assist in improving the performance.

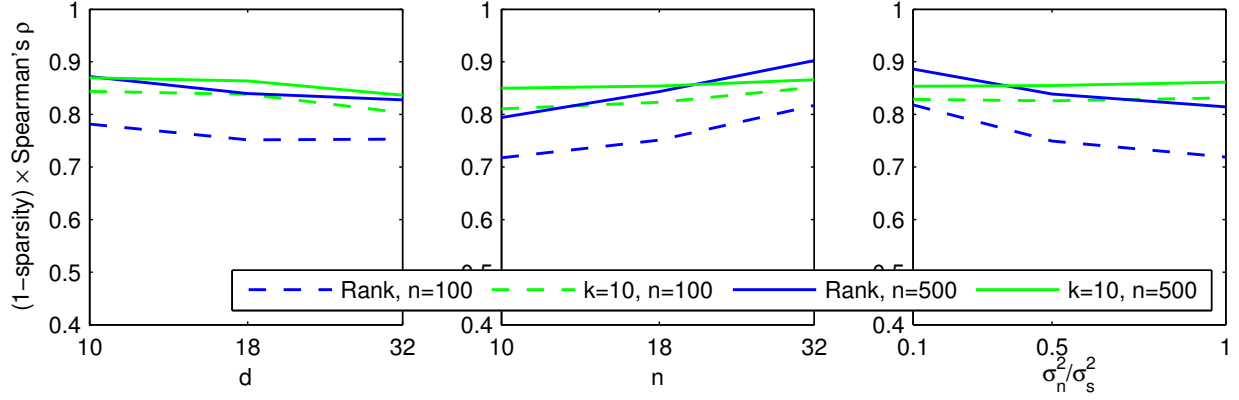


Figure 6: Comparison of retrieval performance between probabilistic metrics estimated by straight-forwardly picking every k -th posterior sample, and estimated after learning a weight vector that preserves ranks. The evaluation metric focuses on whether the method has improved retrieval performance while decreasing the number of posterior samples stored. The experiments are simple regression tasks $\mathbf{w} : \mathbb{R}^d \rightarrow \mathbb{R}$ not forming any clusters. The figures show the variation of performance as a function of the dimensionality (d), number of measurements (n) and signal-to-noise ratio, where the other two features have been averaged over. The number of posterior samples are either 100 or 500. We observe that the proposed approach is generally not better than storing arbitrary posterior samples. However, as more posterior samples are given the proposed approach soon catches up. See section 4.3 for details.

4.2 Comparison between $\widehat{\text{ML}}_{q|d}$ and $\widetilde{\text{ML}}(q|d)$: a representative training set

We use the same dataset to compare performance between the proposed approach for reducing the posterior samples to be stored with weighted average of likelihood $\widehat{\text{ML}}_{q|d}$, and a simpler method. To recapitulate, our goal has been to select from a pool of posterior samples informative ones that can maintain the retrieval performance. Thus the performance should be better than by simply storing every k -th posterior sample without any optimization, $\widetilde{\text{ML}}_{q|d}$. A small k would improve computational time but degrade sparsity. We compare the performance of the proposed approach against performance with $k = 10$ in Fig. 5. Since our objective is to impose sparsity in the weight vector while improving retrieval performance, we evaluate the performance in terms of $(1 - \text{sparsity}) \times \text{minimum-average-precision}$ (see Fig. 2). For the rank preservation approach, we present the result for $K = 25$ since the others $K = 5, 10, 15, 20$ perform equally well. Therefore, we conclude that in the presence of representative experiments in the training set, the proposed approach can safely select informative posterior samples.

4.3 Comparison between $\widehat{\text{ML}}_{q|d}$ and $\widetilde{\text{ML}}(q|d)$: training set not representative

In the previous two sections, we have presented results for the case when the experiments come in clusters, and the query belongs to one of them as well. Intuitively this is a simpler problem since a query always has certain representative experiments in the training set. To elaborate, one can learn the weights for an experiment by preserving ranks within the same cluster, and since the query is from one of the clusters, the learned weight can be used to compute the likelihood of a query reliably. To make the problem difficult we consider the situation when the experiments do not have clustered structure. To investigate if the proposed method still performs well in this ‘extreme’ set-up, we now generate 200 experiments in the same way but without splitting them in clusters. Since now we do not have any ground truth, we consider the ranking given by $\widehat{\text{ML}}_{q|d}$ with all posterior samples as the ground truth, and evaluate the performance in terms of Spearman’s correlation with the ground truth ranking. We observe (Fig. 6) that when the total number of

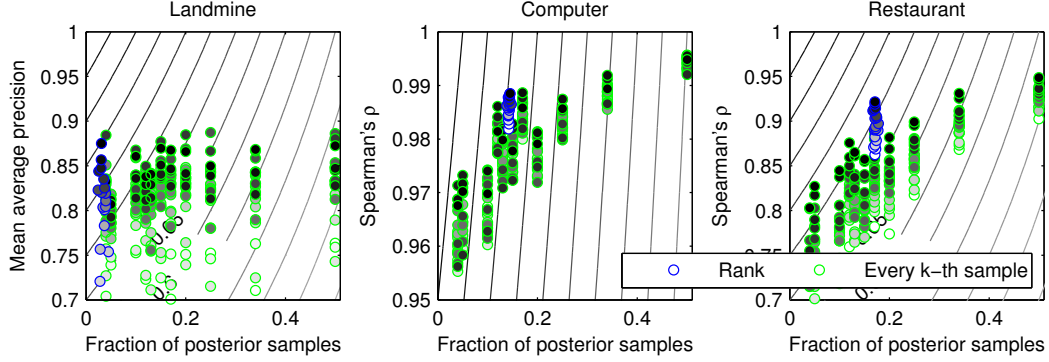


Figure 7: Comparison of the proposed approach and a simpler metric on real datasets. For landmine we present mean average precision MAP as have access to labels of each experiment, while for the other two datasets we present the performance compared to $\widehat{ML}_{q|d}$ estimated with all posterior samples. Each gray shade corresponds to a random partition of the dataset in database and queries. The proposed approach shows improved performance compared to storing every k -th sample since its performance is toward the upper-left corner. The contours are for $(1\text{-sparsity}) \times \text{retrieval-performance}$ (see Fig. 2).

posterior samples m is low, storing every k -th sample performs better. However, as more posterior samples are added, the proposed method performs equally well. This situation is analogous to the general finding that generalizing beyond the learning data is difficult.

5 Experiments

We demonstrate the performance of the proposed approach on three real world datasets: landmine [11], computer [14], and restaurant [15]. The first two are standard in the multi-task learning framework. For landmine, we have access to class labels of each experiment, and we evaluate the performance of our approach in terms of mean average precision MAP, while on the other two datasets we use correlation with respect to the ranking given by $\widehat{ML}_{q|d}$ with all posterior samples. We present the results collectively in Fig. 7. For landmine, we train a binary probit regression model, while for the other datasets we use a normal regression model with non-sparse gamma prior over the weight precisions. For each experiment we generate 100 posterior samples. For each dataset we randomly split it 3:1 into the database and queries.

5.1 Landmine

The data consist of 29 experiments: each experiment is a classification task for detecting the presence of either landmine (1) or clutter (0) from 9 input features. Each experiment has been collected from either a highly foliated region or a desert-like region. Thus they can be split in two classes (16-13). We observe that this is a relatively simple problem in the sense that the classes are well separated, and thus a few posterior samples are sufficient for good retrieval performance. Due to the same reason, the proposed approach is able to retain the retrieval performance using only very few posterior samples.

5.2 Computer

The data consist of 200 experiments: each experiment is a prediction task of how a student rates 20 computers in the scale 0-10. Each computer is described in 13 binary features. Thus, each experiment $\mathbb{R}^{13} \rightarrow \mathbb{R}$ has about 20 samples (some entries missing). Since there are no obvious ground truth labels, we measure how well the proposed approach can reduce the number of posterior samples while preserving rankings. We

observe that the problem is relatively simple since even a few posterior samples have been able to preserve the ranking with respect to $\widehat{\text{ML}}_{q|d}$. However, the number of samples stored is larger than in the previous example since there is no clear clustering.

5.3 Restaurant

The data consist of 119 experiments: each experiment is a prediction task of how a customer rates 130 restaurants in the scale 1-3. All customers do not rate all available restaurants, and so the number of observations in each experiments varies, from 3-18. We select 7 categorical features for each experiment and binarize them, resulting in a $\mathbb{R}^{22} \rightarrow \mathbb{R}$ regression problem. We observe that this problem is more difficult in the sense that performance drops when the number of samples is decreased. However, the proposed approach has been able to collect essential samples to preserve the true rank better.

6 Conclusion and future work

In this paper we have explored the task of retrieving relevant experiments given a query experiment. The state of the art is to retrieve by matching textual (categorical) annotations. We argued that this approach is not optimal since it ignores the actual measurements collected within the experiment, and suggested retrieving experiments based on the relation between covariates and outcomes that is learned from the measurements. However, rather than using a single instance of this relation, we showed that it is better to model its posterior distribution. We used a retrieval metric that computes the marginal likelihood of a query experiment on the models learned on the measurements from existing experiments.

This paper is intended to be a proof of concept towards a potentially highly useful community effort of extending experiment databanks to include also knowledge of the experimenters in a rigorously reusable form, as models. As of now, this is highly non-standard yet would be beneficial since the experimenter alone is best acquainted with his/her measurements and is able to train the most sensible model by incorporating his/her experience as prior knowledge. Storing models of experiments can, however, be cumbersome since most often they are not expressed in an analytic form. A widely applicable alternative is to store samples of the posterior; we suggested approaches to select the most informative posterior samples to store. Notice that posterior samples can be generated also when one has an analytic posterior. We have presented a set of convincing results on simulated data with regression as a task, as well as on standard real datasets.

Acknowledgments

This project is partly supported by the Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN, 251170), and the Aalto University MIDE (Multidisciplinary Institute of Digitalisation and Energy) research programme. The calculations presented above were performed using computer resources within the Aalto University School of Science “Science-IT” project.

References

- [1] Dongliang Ge, Jacques Fellay, Alexander J. Thompson, Jason S. Simon, Kevin V. Shianna, Thomas J. Urban, Erin L. Heinzen, Ping Qiu, Arthur H. Bertelsen, Andrew J. Muir, Mark Sulkowski, John G. McHutchison, and David B. Goldstein. Genetic variation in IL28B predicts hepatitis c treatment-induced viral clearance. *Nature*, 461(7262):399–401, September 2009.
- [2] L. Lukk, M. Kapushesky, J. Nikkilä, H. Parkinson, A. Goncalves, W. Huber, E. Ukkonen, and A. Brazma. A global map of human gene expression. *Nature Biotechnology*, 28(4):322–324, 2010.

- [3] Gabriella Rustici, Nikolay Kolesnikov, Marco Brandizi, Tony Burdett, Mirosław Dylag, Ibrahim Emam, Anna Farne, Emma Hastings, Jon Ison, Maria Keays, Natalja Kurbatova, James Malone, Roby Mani, Annalisa Mupo, Rui Pedro Pereira, Ekaterina Pilicheva, Johan Rung, Anjan Sharma, Y Amy Tang, Tobias Ternent, Andrew Tikhonov, Danielle Welter, Eleanor Williams, Alvis Brazma, Helen Parkinson, and Ugis Sarkans. ArrayExpress update—trends in database growth and links to data analysis tools. *Nucleic acids research*, 41(Database issue):D987–990, January 2013.
- [4] Jr Baumgartner, William A, K Bretonnel Cohen, Lynne M Fox, George Acquaaah-Mensah, and Lawrence Hunter. Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*, 23(13):i41–48, July 2007.
- [5] W.L. Buntine, J. Lofstrom, J. Perkio, S. Perttu, V. Poroshin, T. Silander, H. Tirri, A. Tuominen, and V. Tuulos. A scalable topic-based open source search engine. In *IEEE/WIC/ACM International Conference on Web Intelligence, 2004. WI 2004. Proceedings*, pages 228–234, 2004.
- [6] Christopher J. C. Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Gregory N. Hullender. Learning to rank using gradient descent. In *ICML*, pages 89–96, 2005.
- [7] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June 2008.
- [8] James Malone, Ele Holloway, Tomasz Adamusiak, Misha Kapushesky, Jie Zheng, Nikolay Kolesnikov, Anna Zhukova, Alvis Brazma, and Helen Parkinson. Modeling sample variables with an experimental factor ontology. *Bioinformatics*, 26(8):1112–1118, April 2010.
- [9] Ritabrata Dutta, Sohan Seth, and Samuel Kaski. Retrieval of experiments with sequential dirichlet process mixtures in model space. *arXiv:1310.2125 [cs, stat]*, October 2013.
- [10] Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf. Learning from distributions via support measure machines. *arXiv e-print 1202.6504*, February 2012.
- [11] Y. Xue, X. Liao, L. Carin, and B. Krishnapuram. Multi-task learning for classification with Dirichlet process priors. *Journal of Machine Learning Research*, 8:35–63, 2007.
- [12] José Caldas, Nils Gehlenborg, Ali Faisal, Alvis Brazma, and Samuel Kaski. Probabilistic retrieval and visualization of biologically relevant microarray experiments. *Bioinformatics*, 12(25):i145–153, October 2009.
- [13] Mark D. Smucker, James Allan, and Ben Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, CIKM '07*, pages 623–632. ACM, 2007.
- [14] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 41–48. MIT Press, Cambridge, MA, 2007.
- [15] Blanca Vargas-Govea, Juan Gabriel González-Serna, and Rafael Ponce-Medellín. Effects of relevant contextual features in the performance of a restaurant recommender system. In *Workshop on Context Aware Recommender Systems (CARS)*. 2011.